OPEN VOICE NETWORK

SYNTHETIC VOICE FOR CONTENT OWNERS AND CREATORS

OPEN VOICE NETWORK

Report of the Synthetic Voice Workgroup

March 2022

About Synthetic Voice Considerations for Content Creators and Owners

The emergence of synthetic voice and the spread of voice-controlled platforms, devices, and applications now intersecting and overlapping raises critical questions about the use, value, and ethical impact of synthetic voice. Synthetic voice can be a tool for inclusivity, accessibility, access to expanded knowledge/training, and operational efficiency. It can be misused to entertain or inspire across languages and dialects. It also can be used to manipulate for purposes of commercial fraud or misleading the public for political gain.

We need to ask questions:

What if the voice coming out of the end points is a hyper-realistic representation of a public figure's voice? A famous actor or celebrity? A sports announcer? How about a journalist? Or the president of a public company, or the leader of a country? Does it matter if the person is famous and easily recognized by many? What if the voice is yours or mine? And does it matter if the person is alive or dead, or even how long ago its original owner passed away? Who owns a voice? Does the public need to be made aware that the voice they hear has been synthesized? And not only that it was synthesized, but that it has been done so legally and licensed by its owner for this purpose?

This paper addresses these questions and explores how we might channel this remarkable advancement to societal good and economic advantage. The commercial potential for using synthetic voice is an obvious part of the voice-first digital business models. We explore examples of synthetic voice use cases, from automating and localizing audio books and instructional content to broadcasting and engaging the public at scale. We ask: Do the possibilities outweigh the risks? If not, how can we mitigate the risks? What collectively developed guidelines and standards might be developed to guide this?

Join us in ensuring that synthetic voice works for everyone. Visit openvoicenetwork.org.

Table of Contents

“Words mean more than what is set down on paper. It takes the human voice to infuse them with deeper meaning.”- Maya Angelou

Synthetic Voice Considerations for Content Creators and Owners

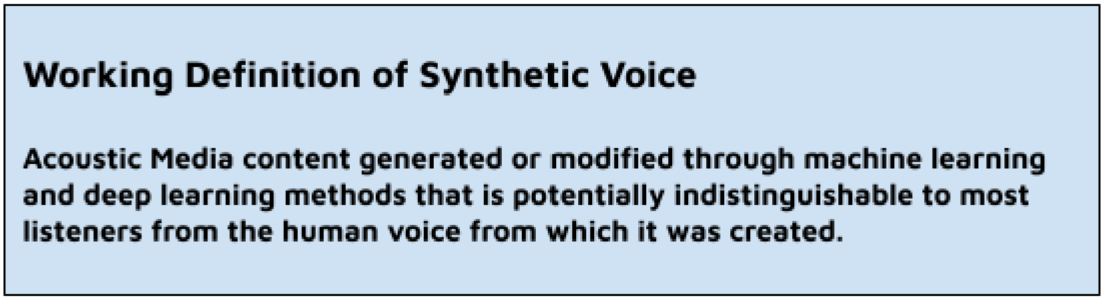
What Is a Synthetic Voice? How Is It Created?

A synthetic voice is a machine-generated replica of a natural human voice that captures the subject’s tone, patterns of speech, and emotional modulation. It is created using text-to-speech technology (TTS) to match the quality of a recorded individual’s voice while generating a convincing copy. Service providers, such as Altered AI, BeyondWords, Respeecher, and Veritone also provide speech-to-speech (STS) processes as well. Typically, speech synthesis is used by developers to create voice robots, such as IVR (Interactive Voice Response). In contrast, STS voice synthesis, powered by AI, allows you to use speech instead of text as a source to generate speech in another voice.

A successful synthetic voice is lifelike because a computer analyzes a person’s voice parameters in many hours of recording, which enables it to model the unique patterns of speech and expression of the person it was built to mimic. Voice quality parameters include vocal resonances, formant dynamics, pitch, loudness, and respiratory dynamics.

Earlier synthetic speech approaches traded off naturalness and intelligibility due to computational limitations. The smallest unit of meaningful sound in a language, such as the “c” in “car,” is a phoneme. One earlier approach relied on phoneme concatenation: slices of recorded speech are “chosen” by the computer from a stored set for best sound in the context of a particular word. Diphone or triphone concatenation used larger segments, each with more natural results. To generate speech, voice assistants typically convert voice commands into text and compare that text to recognizable words in a database. Another early approach, formant synthesis, produced characteristically robotic-sounding speech, relying on acoustic models instead of recorded segments.

Synthetic speech generated by computers has improved markedly since 2017, fueled by significant improvements in speech technology and computing power. We have reached the point at which it can be impossible for human listeners to tell if the **synthetic voice** they hear is generated by AI or recorded by a human. This type of hyper-realistic audio is a subset of computer-generated synthetic media, which includes realistic voices, images, videos, and music**.**



Working Definition of Synthetic Voice

Acoustic Media content generated or modified through machine learning and deep learning methods that is potentially indistinguishable to most listeners from the human voice from which it was created.

Since the audio itself often can’t be distinguished from its individual human originator, synthetic voice raises significant ethical, social, and legal challenges. These challenges are important to address because human voices are part of our human identities. They’re what Henry Wadsworth Longfellow called “the organ of the soul.”

The purpose of this report is to identify and articulate issues surrounding synthetic voice for the intellectual property rights of content creators, which the Open Voice Network respects, and review current/emerging use cases for the legal and ethical practices governing ownership, including explicit transparency and consent, which the Open Voice Network promotes.

Synthetic media first captured broad public attention in 2017 in the form of unauthorized "deepfake" videos on social media that altered the appearance, actions, and statements of people to look and sound like public figures doing and saying embarrassing, misleading, and potentially dangerous things that they would not choose to do or say—an inauspicious beginning with significant intentional and unintentional consequences. When weaponized, deepfakes can inflict harm on individual reputations and societies. However, more productive and exciting use cases of synthetic voice point to it as a scalable, cost-effective way to produce and localize customized digital content at scale.

The Stakeholders

Synthetic voice is worthy of study because of the significant economic, social, and legal implications for the individual and corporate/organization stakeholders who create, own, license, and consume it, including:

- **Audiences, adults and children,** who enjoy being engaged by voice content, and especially in an emerging age of interactive, immersive entertainment.

- **Companies/Organizations** that want synthetic voice created to assist with their endeavors, such as advertising, entertainment, gaming, on-brand voice assistants and messages at scale, search, and other voice-first platform strategies. They seek operational efficiency, extension of a brand or an executive voice throughout all audio communication, including employee education, constituted communication, and connection to constituents through voicebots or voice assistants, and multi-language communication. However, it is critical for these users to know the voicemail they receive is the authentic voice of authority for them. For example:
    - **company spokespersons and executives** who may go on record with news organizations for earnings calls,[1] etc., must ensure their voice's

[1] Veritone Press Release. March 10, 2022. "Veritone Showcases Multilingual Synthetic Voice of Executive Leadership on Veritone's Fourth Quarter and Full Year 2012 Financial Results Call." https://www.veritone.com/press-releases/veritone-showcases-multilingual-synthetic-voices-of-executive-leadership-on-veritones-fourth-quarter-and-full-year-2021-financial-results-call/

authenticity is beyond question. Some find synthetic voices suitable as news readers.[2]

- ○ **Synthetic voice can be a source of regional, location-specific, and diverse** voices for announcements, audio books, and manuals in multiple languages, quickly at reduced cost. This brings data connection to those regardless of their ability to read.

- **Platforms and creators, such as agencies, composers, musicians, producers, podcasters, and internal marketing teams**, to quickly generate audio content without the time and expense of a recording studio. Distribution platforms have been popping up increasingly as business models.

- **Service companies that create synthetic voices with state-of-the-art speech science**, such as Altered AI, Descript, Resemble, Replica Studio, BeyondWords, Respeecher, Speechkit, Veritone, and VocaliD, an early synthetic speech practitioner that creates unique digital voices for individuals who rely on assistive technologies.

- **Voice-over agents, managers, performers who traditionally insure their voices, professional voice-over actors, entertainment lawyers who specialize in contracts, and performers' unions**, such as SAG-AFTRA, estimated to represent 160,000 performers and media professionals in the United States who work in film and digital motion pictures, television programs, commercials, video games, corporate/educational and non-broadcast productions, new media, television and radio news outlets, as well as major label recording artists.

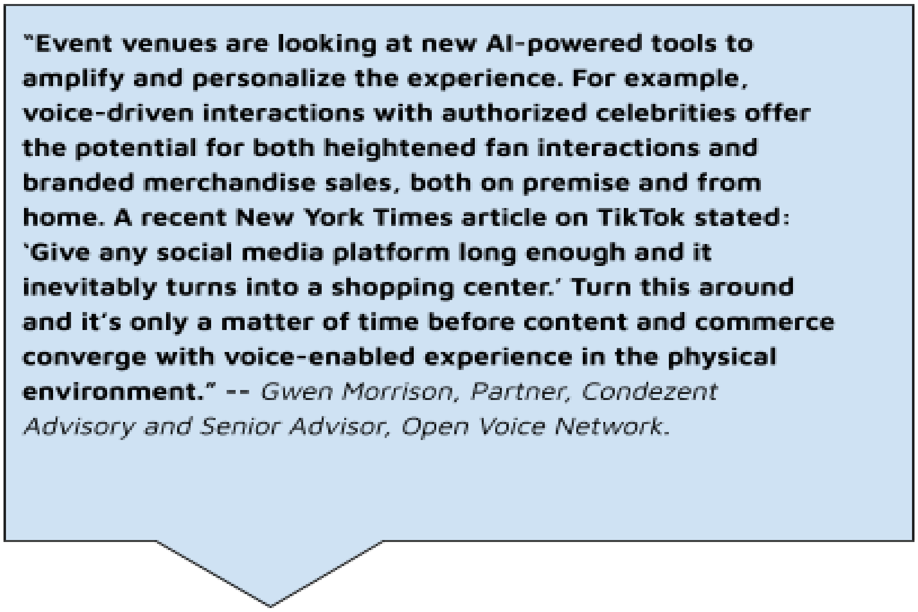
- **Voice-over marketplaces/casting sites**, such as Voice123.com, Bodalgo, CastVoices, and others.

A pioneer of custom-crafted personalized synthetic voices, VocaliD, started out creating unique voices for individuals who rely on speech-assistance devices and have now expanded to create branded voices for businesses and organizations that understand the power of voice for trust, connection, and engagement.

[2] "BBC News Releases Synthetic Voice Newsreader for Online Articles," by Eric Hal Schwartz, *Voicebot.ai,* November 17, 2020. https://voicebot.ai/2020/11/17/bbc-releases-synthetic-voice-newsreader-for-online-articles/

Currently, we are seeing synthetic voice-as-a-service companies emerge and branch out to provide state-of-the-art media creation tools to individuals for personal and commercial use. Companies reported to be engaged in conversation with celebrities and studios to discuss licensing arrangements include Altered AI, Descript, BeyondWords, Resemble, Replica Studio, Respeecher, Speechkit, and Veritone. As an example, Veritone's synthetic voice was born out of the high consumer demand for content and the real struggle for athletes, sports announcers, celebrities, influencers, and the like to meet the needs due to scheduling and other time constraints. There was also the need to bring more authentic and realistic voice content to film and TV dubbing and audio description, audiobooks, podcasts, as well as news, weather, and financial reporting—in multiple languages.

Respeecher uses proprietary deep learning techniques to produce high-quality synthetic speech for Hollywood. Its technology recreated the voice of young Luke Skywalker, which appeared in the final episode of "The Mandalorian," a Star Wars film spinoff. The company also helped to create synthesized speech of former U.S. President Richard M. Nixon for the MIT Open Documentary Lab, which won an Emmy in 2021. Recently, Respeecher has expanded its application to patients with impaired speech to enable them to communicate intelligibly in real time.



"Event venues are looking at new AI-powered tools to amplify and personalize the experience. For example, voice-driven interactions with authorized celebrities offer the potential for both heightened fan interactions and branded merchandise sales, both on premise and from home. A recent New York Times article on TikTok stated: 'Give any social media platform long enough and it inevitably turns into a shopping center.' Turn this around and it's only a matter of time before content and commerce converge with voice-enabled experience in the physical environment." -- *Gwen Morrison, Partner, Condezent Advisory and Senior Advisor, Open Voice Network.*

Use Cases and Potential Benefits

The following use cases for synthetic voice demonstrate its current and anticipated value:

- **Companies, organizations, and educators can use synthetic voice for internal and external functions to**:
  - **Create lifelike, brand-appropriate chatbots** as a lower cost alternative to interacting efficiently with the public
  - **Enable access to multiple brand voices** on a single platform
  - **Translate/dub/hyper-localize content** with diverse voices, regional accents, dialects, or languages cost-effectively
  - **Update existing IVR systems regularly**, and make experiences more dynamic
  - **Speed up design and development efforts** for voice assistants, by allowing quick high quality placeholder creation while the content is still in flux
  - **Archive voices** to collect auditory data.
  - **Preserve indigeneous or diverse-sounding voices**, e.g., in stories that showcase the cadence of how a language flows
  - **Give audio instructions to patients**, particularly engaging, and useful for the functionally illiterate and those without sight
  - **Facilitate advanced manufacturing** with audio manuals, FAQs, and a voice-controlled, touchless manufacturing floor
  - **Aid learning** with engaging audio presentations, exercises, and drills
  - **Access a (potentially) lower cost alternative** in the creation of marketing and promotional materials, which is particularly helpful to young companies
  - **Provide audio access to corporate communications, eLearning courses, and training materials** to engage employees and meet accessibility compliance requirements
  - **Attach a brand's custom voice to pop-up ads** on smart TVs and other streaming devices
  - **Offer monthly subscriptions/fee-based services to create synthetic voices**

    - **Use voice as part of a biometric identification system** with authentication for airports, entrance to corporate campuses.

- **Entertainment and media company creators use synthetic voice to enhance and protect the value of their projects/products and increase their accessibility**:
    - **Film/TV/cartoon creators** can use synthetic voice to:
        - **post-sync movies** in different languages using the voice of the original actor
        - **complete projects** in the event an actor becomes unavailable during or after production—*only with artist’s/artist estate’s permission,* as was the case when the director and crew of Roadrunner: A Film About Anthony Bourdain” claims he asked Bourdain’s widow and literary executor permission to create three soundbites with the celebrity chef’s voice to narrate passages of the film, which his widow disputes[3]
        - **tweak performances and script wording**, in less time without needing the actor physically present. *Careful. Permission required*
        - **monetize/repurpose** existing assets, such as by licensing the voices of cartoon characters for new experiences
        - **resurface the voices of historical figures for a contemporary audience**
        - **create new voice-driven experiences**
    - **Game Developers** use synthetic voice:
        - **as a placeholder** during production for a more immersive creator experience
        - **to modify existing games**, *careful, only with artist permission*
        - **for in-game commerce**, offering opportunities for audio advertising and in mobile gaming so as not to intrude on game play
    - **News organizations** turn text into distinctive audio for expanded distribution and as an alternative way for individuals to engage with the news
    - **Podcasters**:
        - **Localize podcasts** to reach non-english speaking populations

[3] “Anthony Bourdain’s voice was deepfaked in a new film. The chef’s widow and critics aren’t happy.” by Timothy Bella, *The Washington Post*, July 16, 2021

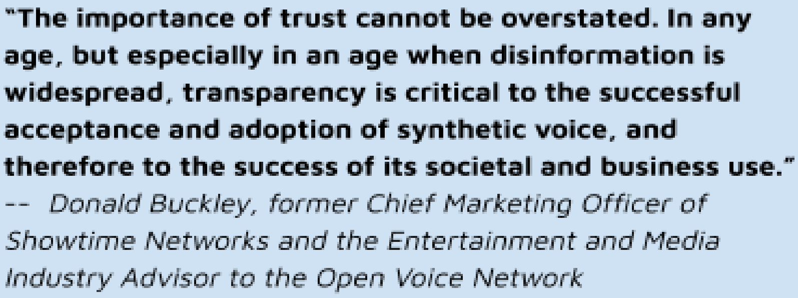
        - **Deliver ideas at scale** with more engaging synthetic personalities
    - **Book and periodical publishers**:
        - **create, translate, and update audio publications** with customized voices faster, cost-effectively, and to make them more widely accessible in the expanding world of digital distribution
        - **have articles read by recognizable voices** that inspire confidence/trust
        - **make specialty/research and academic texts available** via high-quality audio for greater distribution and accessibility
    - **Storytellers/creative inventors** can use synthetic voice to allow a wider universe to bring stories to life via lower-cost "voice actors," not necessarily voices of current working professionals in the field.

- **Professional voice actors capture, mimic, archive synthetic voices to license, expand, market, and protect their revenue stream**:
    - **Recognizable celebrities**, such as Samuel L. Jackson and Morgan Freeman, and popular, trusted voices have tremendous new opportunities to license their synthetic voices for advertising, voice-overs, and customizing experiences, such as authenticating ticket purchases and welcoming audiences
    - **Professional voice-over actors** may archive their voices to expand current and protect opportunities:
        - **Actor Val Kilmer had a synthetic voice created by Sonatic** when he was losing his voice to throat cancer, although it was not used in the documentary for which it was envisioned.
        - **Busy announcers and other voice-over artists** can use synthetic voice to expand their reach and revenue by working on multiple projects simultaneously.

- **Individual consumers can use personalized synthetic voices**:
    - as **assistive technology** to compensate for temporary or permanent voice loss
    - as **a means to grieve** a loved one
    - as **a way to pass on their vocal legacy** to loved ones
    - as **a screen reading and learning aid** for those with visual and/or processing disorders

- for **nostalgia**
- as **a language learning and oral practice tool**.

Potential Harms to Address



"The importance of trust cannot be overstated. In any age, but especially in an age when disinformation is widespread, transparency is critical to the successful acceptance and adoption of synthetic voice, and therefore to the success of its societal and business use."
-- *Donald Buckley, former Chief Marketing Officer of Showtime Networks and the Entertainment and Media Industry Advisor to the Open Voice Network*

Does the public have the right to know if the voice they are hearing is a real person or a synthetic personality voice?

When the global use of synthetic voice begins to mature. Moral consequences tend to be the first line of the public's defense against fast-moving new technology's consequences— both intended and unintended— before legal updates catch up.

Perhaps a cautionary tale is the outcry when it was revealed that 80's R&D duo, Milli Vanilli, had been passing off other voices as their own in their synthesized music and in live concerts. Their contracts specified they would be singing. Fans loved their music, but many resented the lack of transparency about whose voice they were really listening to/contracting to do business with. Milli Vanilli lost their Grammy award and livelihood/creative careers because they did not possess the vocal talent represented as their own. Today, synthetic voices have been created by using an individual's voice without permission, as in the legal case of voice actor Bev Standing v. Tik Tok, and others created to deceive and defraud.

When creating a synthetic voice, among the potential ethical, economic, reputational, social, and legal harms to guard against include:

- **Blurring the line between human talent and digital art for malicious purpose** through lack of transparency about the use of synthetic voice

- **Data breaches** through fraudulent voice identification
    - identity theft to obtain financial/sensitive information, or materials by phone
    - Access to high security clearance areas

- **Manipulation of others**, particularly damaging when it erodes public trust, especially damaging for children

- **Impersonation for the purposes of social bullying/harassment**

- **Spreading misleading information through false content**, particularly when it leads listeners to act on harmful information the speaker did not intend

- **False attribution and illegal impersonation** of voice actors and celebrities. Actor's performances, which include their vocal performances, are protected by copyright.
  However, misuse by impersonators can lead to reputational/economic harm if content is off-brand/offensive or the public simply devalues hearing the actor's real voice because they don't believe it is authentic.

For a professional voice over actor like Bev Standing—who brought the first lawsuit[4] about unauthorized use of a synthetic voice—voice is not only part of her identity, it's also essential to her livelihood. It's the first case, but not likely to be the last.

Well-publicized ethical use of synthetic voices is critical to public appreciation and, ultimately, its successful use.

[4] "Actor sues TikTok for using her voice in viral tool," by Cristina Criddle, BBC News, May 10, 2021

Recommended actions to boost benefits, reduce/mitigate harms

- **Further research is required, specifically in the areas of technology advancements and where they could take synthetic voice in 3-5 years, such as:**
  - recognition-detection technologies
  - current thinking from voice/data security experts
  - emerging enterprise use cases: process integration, value quantification

- **Opportunities for the development of industry standards must be identified. Some ideas from this initial study:**
  - Synthetic voice marking
  - Synthetic voice user recognition-detection
  - Voice actor Intellectual Property (IP) protection, ranging from new industry-agreed contracts, decentralized industry registries of synthetic voice commitments, and
  - user identification, authentication, attestation, authorization

- **A means for the public to detect fakes in addition to watermarks, e.g., authentication technology for consumers on devices.** As Brent Weinstein, Chief Innovation Officer at United Talent Agency said: "As advances in audio AI create new opportunities for talent, there is a need for reliable technology to manage and protect these rights."[5] It is important to provide accessible notification that a voice being used is a synthetic voice made with permission of the actor. The public must maintain the ability to trust what they see and hear. Include a watermark, a form of easily accessible notification to ensure transparency. Preferably it is not an audible interruption of content. Devices themselves could use their lights to provide visual watermarks.

[5] https://www.veritone.com/press-releases/veritone-launches-marvel-ai-a-complete-end-to-end-voice-as-a-service-solution-to-create-and-monetize-hyper-realistic-synthetic-voice-content-at-commercial-scale/

- **Governance, guidelines:** Establishment of rules, working toward something like the VAST template in advertising; establish a body responsible for governance.

- **Broad public education —** to explain the technology, its potential value and benefits, its potential harms, and steps enterprises and individual users should take.

Conclusion

The technologies, uses, and economic value of voice synthesis are in their early days. In a world increasingly shaped by digital content, standards related to synthetic voice use is a topic of importance. In an immersive world of real-time interactive 3D, where lines blur between the real and the created, standards related to synthetic voice are critical.

Let's continue the discussion.

About the Open Voice Network

The Open Voice Network (OVON) is a non-profit industry association dedicated to the development of standards for voice assistance transparency, consent, limited collection and control of voice data that will make using voice technology worthy of user trust. In any reality, virtual or otherwise, we believe personal privacy should be respected as the default. The Open Voice Network operates as an open-source community within The Linux Foundation. It is independently funded and governed with participation from more than 120 voice practitioners and enterprise leaders from 12 countries.

The Open Voice Network community's work is open source. We seek inclusive input and like to share our insights. At present, our work is focused in four areas:

- **Interoperability**, defined as the ability for conversational agents to share dialogs (and accompanying context, control, and privacy),

- **Destination registration and management**, the ability of users to confidently find a destination of choice through specific requests, and for the providers of goods and services to register a verbal “brand” — similar to the Domain Name System (DNS) of the internet;
- **Privacy**, with voice-specific guidance for both the protection of individual user data and that of commercial users; and
- **Security**, with a focus on voice-specific threats and harms.

Please see our papers in 2022 and support the Open Voice Network by visiting openvoicenetwork.org.

About the Linux Foundation

Founded in 2000, The Linux Foundation is supported by more than 1,000 members and is the world’s leading home for collaboration on open-source software, open standards, open data, and open hardware. The Linux Foundation’s projects are critical to the world’s infrastructure including Linux, Kubernetes, Node.js, and more. The Linux Foundation’s methodology focuses on leveraging best practices and addressing the needs of contributors, users and solution providers to create sustainable models for open collaboration. For more information, please visit us at linuxfoundation.org.

Acknowledgements

Authored by the Open Voice Network with special thanks to Synthetic Voice Workgroup contributors, Audrey Arbeeny, Ashley Bailey, Donald Buckley, Anna Bulakh, Jacqueline Carly, Oita Coleman, Suze Cooper, Lauren Cowles, Colleen Fahey, Judy Fossum, Alyssa Frank, Anne Ganguzza, Ann Thyme-Gobbel, David James, Juliet Jordan, Jim Kennelly, Sean King, Jodi Krangle, Adam Lofbomm, James MacLeod, Janice Mandel (draft writer), Andy Maskin, Patrick O'Flaherty, Rupal Patel, Alex Serdiuk, Jon Stine, and Rich Sutton.

Licensing and Attribution

Happy to share our work. Before you cite it, here's what you need to know:

This Open Voice Network white paper has been developed within the Vocabulary and Definition Project of the Open Voice Network, which has been established as Vocabulary and Definition Project a Series of LF Projects, LLC (the "Project"). LF Projects, LLC ("LF Projects") is a Delaware series limited liability company.

LF Projects hold title to all trade or service marks used by the Project ("Project Trademarks"), whether based on common law or registered rights. Project Trademarks will be transferred and assigned to LF Projects to hold on behalf of the Project. Any use of any Project Trademarks by Collaborators in the Project will be in accordance with the license from LF Projects and inure to the benefit of LF Projects.